Panel Submission

Title: Explainable Intelligent Systems and the Trustworthiness of Artificial Experts

Chair: --

Group: Research Project "Explainable Intelligent Systems (EIS)"

Presenters:

Kevin Baum (Philosophy/Computer Science, Saarland University), Timo Speith (Philosophy, Saarland University), "Why Explainable AI Matters Morally"

Felix Bräuer (Philosophy, Saarland University), Eva Schmidt (Philosophy, University of Zurich), "Artificial Intelligent Systems: Reasonable Trust Requires Rationalizing Explanations"

Lena Kästner (Philosophy, Saarland University), Daniel Oster (Philosophy, Saarland University), Andreas Sesing (Legal informatics, Saarland University) "What Kind of Explanation Shall Artificial Experts Provide?"

Tina Feldkamp (Psychology, Saarland University), Markus Langer (Psychology, Saarland University), "How the Kind of Explanation Affects People's Reactions to Artificial Experts"

Organizational note: Although each paper has several authors, it will be presented by only one person at the panel.

Abstract Overview:

We frequently rely on experts and their judgments: We consult a doctor to check whether we need additional vaccines for our trip to Asia, check the weather forecast to pack for our trip, use a navigation app to get around unknown cities, ... Some of the experts we rely on are humans. But increasingly, we rely on algorithms instead. Algorithms do trivial things such as recommending movies on Netflix or books on Amazon. But they are also used in high-stakes situations, e.g. to determine whether an applicant should receive a loan. Similarly, judges rely on algorithms to decide whether to grant parole (Corbett-Davies et al. 2017) and doctors employ algorithms to determine whether a patient should be taken off life-support (Song et al. 2018) or receive palliative care (Avati et al. 2018).

In this interdisciplinary panel, we bring together experts from philosophy, psychology, law, and computer science to discuss both the prospects and the challenges of relying on artificial experts. The first talk examines why it is morally important for artificial experts to be explainable. The second talk argues that the outputs of trustworthy artificial experts

must be explainable via rationalizing explanations. The third talk relies on specific frameworks of scientific explanation to investigate the role that causal relations may play in explanations of artificial experts' judgments. Finally, the fourth talk critically discusses whether more detailed explanations (of the recommendations provided by artificial experts) guarantee increased user acceptance.

A/V Requirements: blackboard or whiteboard; projector

Individual Abstracts:

1. Why Explainable AI Matters Morally

Kevin Baum (Philosophy/Computer Science, Saarland University), Timo Speith (Philosophy, Saarland University)

We argue that opaque artificial experts threaten to violate the moral rights of those being evaluated. To avoid this, it's not enough to put a human in the loop, because even human experts should only consult artificial experts which they are epistemically justified to trust. We use the real-life example of Loomis v. Wisconsin in order to strengthen our point (cf. Moore 2017). We argue that it is morally permissible to use artificial experts if their recommendations are rationally comprehensible to human decision-makers in a way that allows them to rule out relevant alternatives and to access the reasons for the artificial expert's recommendations. Furthermore, we present our recent research in computer science as a first practical approach. It combines Abstract Dialectical Frameworks (Brewka and Gordon 2010) and practical reasoning in order to develop a well-suited data structure and so delivers the relevant kind of explanation (Baum et al. 2019).

2. Artificial Intelligent Systems: Reasonable Trust Requires Rationalizing Explanations

Eva Schmidt (Philosophy, Zurich), Felix Bräuer (Philosophy, Saarland University)

We argue that autonomous artificial intelligent systems, including artificial experts, need to be explainable by appeal to rationalizing explanations (Baum et al. 2017): Agents ought to use such systems only if they can trust them for good reasons, but they can only do so if it's epistemically accessible to them that the systems are trustworthy, or that trust in them is appropriate. After discussing different levels of appropriate trust, we argue that the right

level of trust to extend to autonomous AI systems is goal-relative trust, which requires that the trustor be able to recognize that the trustee's goal harmonizes with the trustor's goals as well as that the trustee competently pursues the goal. Users, then, need to be able to recognize the goals of such systems and the information they have to go on in pursuit of these goals – they need rationalizing explanations of the systems' behavior (Davidson 1963).

3. What Kind of Explanation Shall Artificial Experts Provide?

Lena Kästner (Philosophy, Saarland University), Daniel Oster (Philosophy, Saarland University), Andreas Sesing (Legal informatics, Saarland University)

To ensure the trustworthiness and fairness of artificial experts, we demand that their judgments be *explainable*. But what does this mean? A common assumption is that explanations describe causes (e.g. Salmon 1998). While there undoubtedly is an important link between explanation and causation, recent discussions in philosophy of science also focus on underlying mechanisms (e.g., Craver 2007, Glennan 2017), difference-making (e.g. Woodward 2003), and network models (e.g. Borsboom et al. 2018). In this talk, we examine to what extent artificial experts' judgments may be explained in terms of these frameworks and what role causation plays in each case. We specifically focus on the question of how to identify causal relations among other explanatorily relevant relations. To achieve this, we combine philosophical accounts of causation with criteria commonly used in legal contexts (penal law, civil law) to constrain liability in problematic cases.

4. How the Kind of Explanation Affects People's Reactions to Artificial Experts

Markus Langer (Psychology, Saarland University), Tina Feldkamp (Psychology, Saarland University)

Algorithms increasingly support decisions in ethically sensitive domains. Therefore, user acceptance of these algorithms becomes increasingly important. To ensure the acceptance (e.g. perceived fairness) of algorithm-based recommendations, it is plausible to require explanations of how relevant algorithms reach their recommendations. It remains unclear, however, which kinds of explanations best serve this purpose (cf. Langer et al, 2018). In a study concerning personnel selection, we experimentally examined the impact of different kinds of explanation on user acceptance. In a 2 (no process information vs. process information) × 2 (no process justification vs. process justification) design, participants (N = 124) re-

ceived explanations and watched a video showing an algorithm-based interview. The results indicate that process justification is better than process information to improve user acceptance (i.e., higher perceived fairness). However, receiving no explanation led to results regarding user acceptance similar to receiving process justification, indicating a complex relation between acceptance and explanations.

Biographies:

Kevin Baum (Philosophy/Computer Science, Saarland University)

Kevin Baum is a PhD student, lecturer and research assistant at Saarland University. He holds two master's degrees (in Philosophy and Computer Science) and his thesis in Computer Science is about foundations of Machine Explainability and Justifiability. Kevin teaches and does research on Computer Ethics, Machine Explainability & Ethics, Normative Ethics and Decision Theory. He is part of Explainable Intelligent Systems.

Felix Bräuer (Philosophy, Saarland University)

Felix Bräuer is an assistant professor of philosophy at Saarland University. Before joining the philosophy department at Saarland University, he wrote a PhD thesis on the epistemology of testimony at the Humboldt University of Berlin. He mostly works in epistemology, especially social epistemology, as well as in philosophy of language.

Tina Feldkamp (Psychology, Saarland University)

Tina Feldkamp is PhD student and research assistant with the industrial and organizational psychology group at Saarland University. Her research interests are on perceived fairness, trust and attribution of responsibility when using algorithm-based systems in Human Resource Management with a focus on comprehensibility of the systems. She teaches on personnel selection and is part of Explainable Intelligent Systems.

Lena Kästner (Philosophy, Saarland University)

Lena Kästner is a junior professor in philosophy of mind and cognitive systems at Saarland University, Germany. Her primary research areas are philosophy of mind and philosophy of science, especially philosophy of cognitive science. Her background is in cognitive science and cognitive neuroscience. She specializes in scientific explanations (particularly explana-

tions of cognitive phenomena), cognitive architectures, experiments, and causation.

Markus Langer (Psychology, Saarland University)

Markus Langer is a Post-Doc with the industrial and organizational psychology group at Saarland University. He wrote his dissertation on novel technologies for job interviews. His research interests cover personnel selection, innovative technologies for human resource management and the relation of humans and artificial intelligence with a focus on user acceptance and comprehensibility of algorithm-based systems.

Daniel Oster (Philosophy, Saarland University)

Daniel Oster (B. A.) is currently finishing his master thesis in Philosophy about Machine Explainability at Saarland University. He examines the requirements on explanations of artificial systems' decisions with special respect to appropriate elucidations for affected people. He is part of Explainable Intelligent Systems, a research assistant in Analytical Philosophy and contributes to the Computer Science lecture "Ethics for Nerds".

Eva Schmidt (Philosophy, University of Zurich)

Eva Schmidt is a researcher at philosophy department at the University of Zurich, where she works in the project *The Structure and Development of Understanding Actions and Reasons*. She is also interested in epistemic reasons. Previously, she worked at Saarland University, where she defended her dissertation on the nonconceptual content of perceptual experience in 2011.

Andreas Sesing (Legal informatics, Saarland University)

Andreas Sesing is a research assistant at Saarland University, chair of Professor Dr. Georg Borges (Faculty of Law), with a focus on IP and IT Law. Law studies at Ruhr-Universität Bochum (2005-2010), first state examination in April 2010. Research assistant at Ruhr-Universität Bochum (2010-2012), legal clerkship at the Landgericht Essen (2013-2015), second state examination in May 2015.

Timo Speith (Philosophy, Saarland University)

Timo Speith is a PhD student in Philosophy and research assistant at Saarland University. He holds a bachelor's degree in Philosophy and a master's degree in Computer Science. His the-

sis investigates which types of explanations are fruitful to explain AI systems. Timo teaches and does researches on Computer Ethics, Machine Explainability and in Philosophy of Sciene especially on Scientific Explanations.

References

Avati, A. et al. (2018). Improving palliative care with deep learning. *BMC medical informatics and decision making*, *18* (4), 122.

Baum, K., Köhl, M. and Schmidt, E. (2017). Two Challenges for CI Trustworthiness and How to Address Them. *ACL Anthology*, 1–5.

Baum, K., Hermanns, H. and Speith, T. (2019). Towards a Framework Combining Machine Ethics and Machine Explainability. arXiv preprint arXiv:1901.00590.

Borsboom, D., Cramer, A.O.J. & Kalis, A. (2018). Brain disorders? Not really... Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*.

Brewka, G. & Woltran, S. (2010). Abstract dialectical frameworks. In Proc. KR'10, 102–111.

Corbett-Davies, S. et al. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806). ACM.

Craver, C. (2007). *Explaining the Brain*. Oxford University Press.

Davidson, D (1963). Actions, reasons, and causes. *Journal of Philosophy* 60, 685–700.

Glennan, S. (2017). The New Mechanical Philosophy. Oxford: Oxford University Press.

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior*, *81*, 19-30. doi:10.1016/j.chb.2017.11.036

Moore, T. R. (2017). Trade Secrets and Algorithms as Barriers to Social Justice. Resource of the Center for Democracy & Technology. URL: https://cdt.org/files/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf.

Salmon, W. (1998). *Causality and Explanation*. Oxford University Press.

Song, M. et al. (2018). Prognostication of chronic disorders of consciousness using brain functional networks and clinical characteristics. *ELife*, *7*, e36173.

Woodward, J. (2003). *Making Things Happen*. Oxford University Press.